

审阅《城市交通事故的局部相关性可视分析方法》有感

《城市交通事故的局部相关性可视分析方法》为刘新月同学和罗老师投稿到图学学报的一篇文章。该文的审稿人陈为老师要求我审阅这篇论文，顾得到了观看这篇论文的机会，从而有了太多的感想。

该文虽没有太高深的算法和逻辑，但我却从该文中读到了做科研的方法。原因我想要以下几点：

1. 论文背景及其熟悉
 - a) 论文中所使用的数据，自己正在使用
 - b) 论文所研究的问题，罗老师也同我交流过。例如局部性特征分析，17年申请浙江大学开放课题的时候就是以考虑交通事故的局部时空模式为研究问题
2. 做论文的过程没有参与
 - a) 不清楚该论文的成行过程，自己也想过由该出发点，如何写论文但思想混乱，不知道如何着笔。
 - b) 该论文的逻辑清晰，能理顺从前思考的混乱思路，让我了解科研的思考路径

下面详细说下论文的逻辑及层层递进的关系：

1. 论文立题（论文书写的基石，后文的论文都是以此为基础，若该论断不成立，则论文将没有意义）：事故在全球范围上呈现随机分布的态势，但在局部范围内会呈现出一定的规律。本文采用事故发生根据天气在全球上的分布比，和个别路段分布比作了对比，从而得出了该立题的正确性。（论文的立题要有证明，或者有其他论文背书）
2. 何为“局部”：论文讨论的是局部相关性，那么就需要明确何为局部，对局部进行限定。论文中的局部被表示为事故多发路段，并对事故多发路段进行了明确的定义。在我考虑事故多发路段的时候，经常会考虑事故多发路段会不会和时间有关系，从而导致陷入定义事故多发路段的迷茫，但文章中将时间作为事故多发路段的一个属性，从而明确了事故多发路段。（定义明确，科研是一个明确的概念，一就是一，二就是二，一定要明确，有清晰的认识，不能模糊，含糊其词）
3. “局部”进阶：只考虑单一的事故多发路段作为局部并没有好的研究手段与研究意义，因此本文，对事故多发路段进行了聚类。以聚类后的对象做为局部。事故多发路段是在地里意义上的聚合，而事故多发路段的聚类则抛开了地里意义上的距离度量，定义了新的距离公式。（如何将问题升级，从而引出合理的解释）
4. 可视化的新意：本文提出了类簇的可视化编码，构建了自己创建的图形。（其实类簇就是直方图的升级版，但经过该图形概念的提出，是文章的可视化贡献提高。这一点是我需要及其注意的，自己在做可视化系统的时候，使用的都是及其简单的可视化图形，要注意对使用图形的整合与改进）

论文的感悟：

1. 论文的解决问题要**明确**。该论文的解决问题就是探索局部相关性
2. 论文的研究点要**小**，最好能用一句话来概括。
3. 科研过程中，要勇于舍弃和敢于创新。或者说是要经常放弃宽度而加高深度。从前我做事的方法是从A直接到B，但科研的方法是从A到B的中间截取2个点如C和D，然后将C到D直接的深度做下去，使得由C到D会取得很好的效果。



图 1 工程科研对比

4. 论文的逻辑要清晰，关系要层层递进有清晰的逻辑结构。

综上，在针对于我的 ChinaVis 论文根据以上几点进行分析。

1. 论文立题（交通事故的发生风险是可以预测）：众多的交通事故预测的论文证明该立题是有一定意义的，但所有论文的预测效果都不太精确也表明这一问题的复杂性。（论文的预测结果都没有采用准确率来评价，而是使用的距离发生数目的差值即 MSE，而该指标的预测结果在事故预测中并没有实际意义）
事故原因最主要是人的因素，缺少驾驶人的状态数据导致事故的预测不会有很强的可靠性。综上，交通事故的发生风险是可以预测的，但是数据的缺失导致该问题暂时不会有很好的解决方案。
交通事故的预测太依赖数据，并不适合我作为研究内容（获取数据十分困难，而且现在的技术手段也不可能获取驾驶员的精神状态数据）。
因此，事故预测也许可以出一两篇论文，但没有研究价值。
2. ChinaVis 论文的选点主要有 2 个。
 - a) 构建深度学习预测框架
 - b) 对深度学习框架加入可视分析模块
 - i. 交互的方式加入用户选择偏好
 - ii. 可视化的呈现预测效果，用户分析出预测不好的路段重新预测

从上可知，**论文需要研究的点太多。**

要构建深度学习框架，就单独构建深度学习框架，并用该深度学习框架与其他的深度学习框架做对比，表明框架的优势。如果做可视分析，就要单一做可视化分析。如果辅助深度学习方法，最好选用一篇已经发表的深度学习框架，并用可视分析的方法对其改进。

因为论文在以上两方面都要写，导致两方面都没有叙述清楚，两方面都缺少必要的细节信息。

可视化功能点过多，减少可视化功能点，把一个功能点介绍清楚。

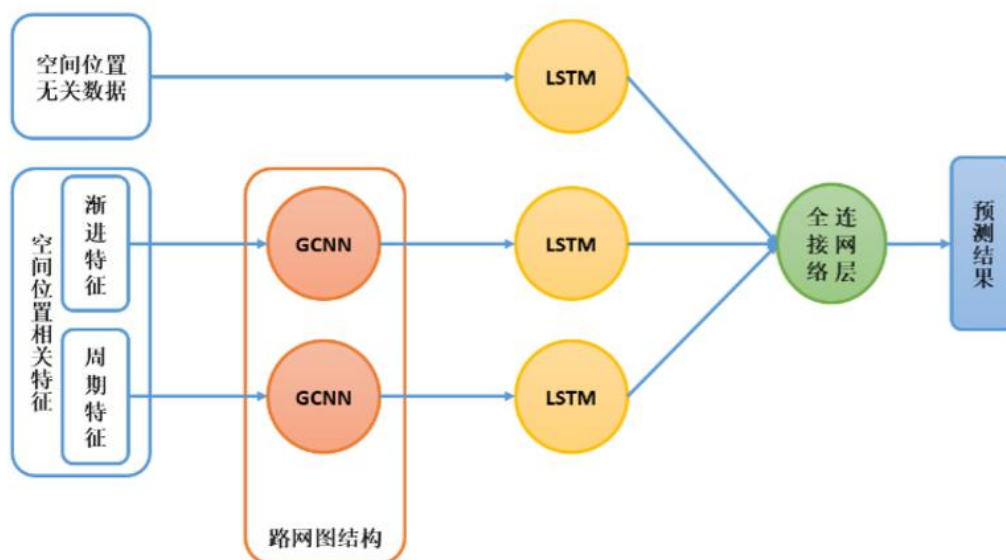
3. 论文的后续改进：

因为还没有使用路段进行事故预测的深度学习模型，所以论文中还需要构建深度学习框架，但是要加入深度学习框架的细节信息。然后将可视化部分的功能精简。用户道路选择偏好，因为没有反馈的部分，因此可以**去掉用户道路选择偏好可视化的贡献点**。可视化的重要贡献在预测结果的可视呈现，用户选出预测不好的路段，重新预测。
4. 论文的投稿规划：论文投稿到 CCF 第七届大数据学术会议。会议的截稿时间由 6 月 15 日调整到 7 月 15 日。因此，继续修改论文，提高论文质量，**保证该论文的录取。**

规划日程安排

前期工作：杭州数据预处理完毕，并用深度学习模型进行了预测，预测准确度在 60%左右。

深度学习模型的参数：



1. 渐进特征的特征输入个数：例如：由前 10 个时间段预测，还是由前 20 个时间预测。
2. 周期特征的采样：预测事故发生，根据交通事故的发生周期性，可以间隔一段时间进行数据采样，然后构建输入序列。这里的采样间隔，不同的路段会有不同的数值，针对预测失真路段可重新选择采样间隔，作为参数。
3. 加强预测失败路段，事故发生记录的训练数目，减少成功路段事故记录的数目

攻克难点：预测效果的可视化呈现。（可视化功能点要单一，但呈现方式要有新意）

1. 构建评价指标计算公式（和罗老师的聚类度量公式类似，由直方图进行聚类）
2. 设计新颖的可视化图形呈现可视化指标
3. 建立良好的交互操作

论文改进点：由合肥单一实例，加入杭州实例。把可视化点做小，做新。

后续的时间安排：

1. 6 月 30 号之前，完成预测指标评价公式的制定，设计出新意的交互式图形
2. 7 月 5 号之前，自己完成对论文的修改
3. 7 月 5 号-7 月 8 号，由罗老师和陈老师和其他审阅修改后的论文。
4. 7 月 9 号-7 月 10 号，根据修改意见修改论文。
5. 7 月 10 号-15 号，（待定，若 10 号把中文论文修改完毕，则这段时间可对论文进行翻译，投稿英文）

全力修改论文，将论文做到最好，争取在会议上的录用（**提高自信心**）。

7 月 15 号之后，考虑一个新的研究问题。（可以是交通方面，但不在做事故预测方面的问题。原因：交通事故的发生风险是可以预测这一立题是有问题的）